A workflow to Discover the variant on Cavities Surface of Proteins

Student: Huimin Lu (Email: luhuimin13@outlook.com)
Supervisor: Mauno Vihinen
Credit: 30 (20 weeks)

Abstract

Protein cavities are a type of protein structure feature. These structures resemble voids inside proteins and can either connect to the outside through a tunnel or remain completely closed. This kind of special protein structure enables proteins to have significant biological functions, such as acting as binding sites for enzymes or facilitating molecule transportation. Furthermore, variants in protein cavities might influence their biological function. However, our understanding of protein cavities and the regulation of variants within them remains unclear.

To explore the features of protein cavities and the variants within them, we employed *CICLOP*, a cavities-detecting software, to identify cavities inside our sample proteins and discover their characteristics. The sample consisted of proteins that contain amino acid substitutions, downloaded from *VariBench*, a benchmark database that includes couple of experimentally validated protein variation datasets. Additionally, we mapped the variants onto the sample proteins to identify mutations located within the protein cavities detected by CICLOP. This allowed us to gain insight into the features of these variants situated on the surface of protein cavities.
The project comprises a workflow that includes several programs. The code and other files can be found in GitHub: https://github.com/luhuim/Variant_in_cavity/tree/master.

Introduction

Background
The functionality of protein depends on protein topological structural features. Cavity is one of structural feature, which is a special class of voids that may be open from one or both ends, or completely enclosed (Garg et al. n.d.).

Cavities in protein structures often serve as sites for enzyme catalysis (Choi and Lee 2019) and tunnels for molecule transportation (Zhou and McCammon 2010). Additionally, inner cavities within proteins can create niche environments that may be associated with protein stability (Tanwar et al. 2013, Xu et al. 2018).

Several cavity-detecting methods have been developed, such as CAVITI (Xu et al. 2018) and MOLEonline (Pravda et al. 2018) etc. However, these tools have some shortcomings. Some methods could only identify cavities by comparing input protein structures with a cavities database, potentially missing unknown cavities (Xu et al. 2018). Furthermore, certain methods may not be fully automated, requiring users to manually select a residue located on the cavity surface (Pravda et al. 2018).

However, CICLOP might address the limitations of previous cavity-detecting methods (Garg et al. n.d.). CICLOP employs a robust grid-based computation method to scan every continuous empty region and effectively identify inner cavities within proteins. This approach can detect cavities that have not been annotated in a database and provide more accurate results without the need for manual settings. And CICLOP can process any protein structure that has PDB file (Garg et al. n.d.; Pravda et al. 2018).

Objective

The sample for this project is a dataset of variant proteins obtained from Varibench (Nair and Vihinen 2013). The dataset primarily consisted of Uniport IDs of variant proteins, the positions of mutations, and the corresponding amino acid substitutions. One of objective of this project is to detect cavities within the variant proteins in sample and uncover their characteristics. Another goal in this project is to filter the amino acid substitutions located within these cavities and investigate the feature of mutations.

These variant proteins should first be mapped to SIFT database to obtain all corresponding PDB structures that include Uniport IDs of variant proteins (Dana et al. 2019). Subsequently, CICLOP is utilized to detect cavities within variant proteins. After cavity identification, an investigation of delineation between cavities and rest part of proteins would be conducted, by analyzing amino acid distribution of protein cavities.

Additionally, the amino acid substitution will be mapped onto the previously identified PDB structures, and the mutations situated on the surface of protein cavities would be left. The amino acid distribution of variants within cavities and the tendency of amino acid substitution will be analyzed.

Furthermore, the Uniport IDs of variants proteins will be converted into gene entry names for Gene Ontology (GO) enrichment analysis (Bateman et al. 2023). And there would be a GO enriched terms comparison between variant proteins in sample and variant proteins that include mutations within cavities.

Method and Material

Material

The sample of this project is amino acid substitution variation data downloaded from *Varibench,* a benchmark database that consist of variation dataset that could be used as ground truth in prediction variation effect. And the datasets are all from experimentally validated human amino acid substitution

variants data and other database (Nair and Vihinen 2013). The dataset is used as training set for PON-ALL model(Yang, Shao, and Vihinen 2022) which is a variant pathogenicity/tolerance predictor for amino acid substitutions in any organism. The sample dataset primarily consisted of Uniport IDs of variant proteins, the positions of mutations, and the corresponding amino acid substitutions.

Method

Data preparation

We firstly find the all PDB structures that contain chains whose Uniport ID mentioned in variant dataset. *SIFT* include information of all PDB structure, including the Uniport ID of each amino acid chain inside PDB structures. After that, we will only leave the variants whose Uniport ID has PDB structures. Then we will go through entities of each PDB structure, and remove the identical chains recorded by *SIFT* and add annotation for each entity (Dana et al. 2019) .

Identifying and discovering the thickness of cavities surface

Then we will download the PDB file and fasta file of each PDB structure that found previously, and detect cavities through CICLOP. We will get the amino acid distribution for protein cavities detected by CICLOP, and check whether this distribution is significantly different from amino acid distribution of entities mapped from variant proteins.

We will firstly compute the proportion of each amino acid present within both entities chosen previously and protein cavities. In this way, we will get two groups of data: one is proportion distribution of the whole entities, another is proportion status of protein cavities. Both groups have same subjects which are twenty kinds of amino acids, and the values are paired according to subjects. Subsequently, a paired t-test will be performed to ascertain whether the mean difference between two sets of percentages is zero. If the p-value was higher than 0.05, than would indicate the amino acid distribution of the whole entities selected previously might not be different from the distribution of protein cavities significantly.

Also, *Wilcoxon signed rank exact test* is non-parametric version of the *paired t-test*, and it can be utilized as supplement test to compare the difference between proportion distribution of the whole entities and proportion status of protein cavities. And if the p-value of *Wilcoxon signed rank exact test* was higher than 0.05, it would be another evidence that the amino acid distribution of the whole entities selected previously might not be different from the distribution of protein cavities significantly.

For every PDB file processed through *CICLOP*, if there were any cavities identified in this PDB structure, there would be a new PDB files generated by CICLOP. This new-generated PDB file includes marks added by CICLOP, and these marks indicate atoms in amino acids that are identified as cavities. In this way, cavities region detected by CICLOP can be regarded as shell-resembled structures that consist of numerous atoms belonging to different amino acids.

To investigate the amounts of atoms with cavity mark that belong to a single amino acid located on the

surface of cavities, we manually set some rules to do further discovery. For example, we intend to know the number of amino acids on cavity surface which only include one atom identified as cavity component. In this case, we set a threshold as 2, which means among amino acids detected on cavity surface only the amino acids that include at least 2 atoms marked by *CICLOP* can be counted as protein cavities, and obtain the difference of amino acids amount between this pseudo protein cavities and cavities detected by *CICLOP*. In this project, we only set the "threshold" as 2 and 3.

Obtaining variant in cavities and discover the bias

We will map the variant with PDB structure, and get the variants that locate in cavities. Then we get the distribution of each amino acid variant, and using hypergeometric function to discover the bias of variant. The function (Figure 1) stimulates non-return sampling, the population is all amino acids in protein cavities, and the draw amount is the number of all substitution amino acids in cavities. For one kind of amino acids, the sampling component can be divided into two parts: this kind of amino acids, regarded as "success" and other 19 kinds of amino acids, regarded as "failure". In this case, we will calculate the probability that the amount of "success" is equal to the amount of this kind of amino acid in variant within cavities. In this way we will get the probability of 20 kinds of amino acids, and the lower probability indicate the more bias. To quantify bias, we introduce fold changed value, and higher fold change value might indicate higher bias. The direction of bias can be defined by *Expected Amount* of one kind of amino acid and the real amount of this type of amino acids. And the *Expected Amount* is calculated from the percentage of this sort of amino acid among the population.

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

N: Sum of amino acids in cavities
K: The amount of one type of amino acid in cavities
n: Sum of amino acids in variant
k: The amount of one type of amino acid in variants

$$Expected\ Amount = \frac{n * K}{N}$$

$k = Expected\ Amount\ (Matched)$:

$$Fold\ Change = 1$$

$k < Expected\ \ Amount\ (Under - Representation)$:

$$Fold\ Change = \frac{Expected\ Amount}{k}$$

$k > Expected\ Amount\ (Over - Representation)$:

$$Fold\ Change = \frac{k}{Expected\ Amount}$$

Figure 1. This figure displays the function of hypergeometric to calculate fold changed value.

Gene Ontology enrichment analysis

Gene Ontology knowledgebase provides controlled vocabulary to describe gene products in three categories, namely biological process, molecular function and cellular component (Ashburner et al.

2000), and GO enrichment analysis is to investigate the functional relevance among given gene sets. To get a summary of function of variant proteins, we convert Uniport ID of variants protein into gene entry IDs Bateman et al. 2023), and obtain overrepresented GO term through GO enrichment analysis, utilizing whole human genome as background. Also, another GO enrichment analysis should be conducted for proteins that have variants in cavities, with background of all variant proteins in sample. The result of two GO enrichment analysis will be compared to discover the functional feature for proteins that have variants within cavities.

Result

Workflow Summary

The whole process of this project could be summarized into a workflow (Figure 2).

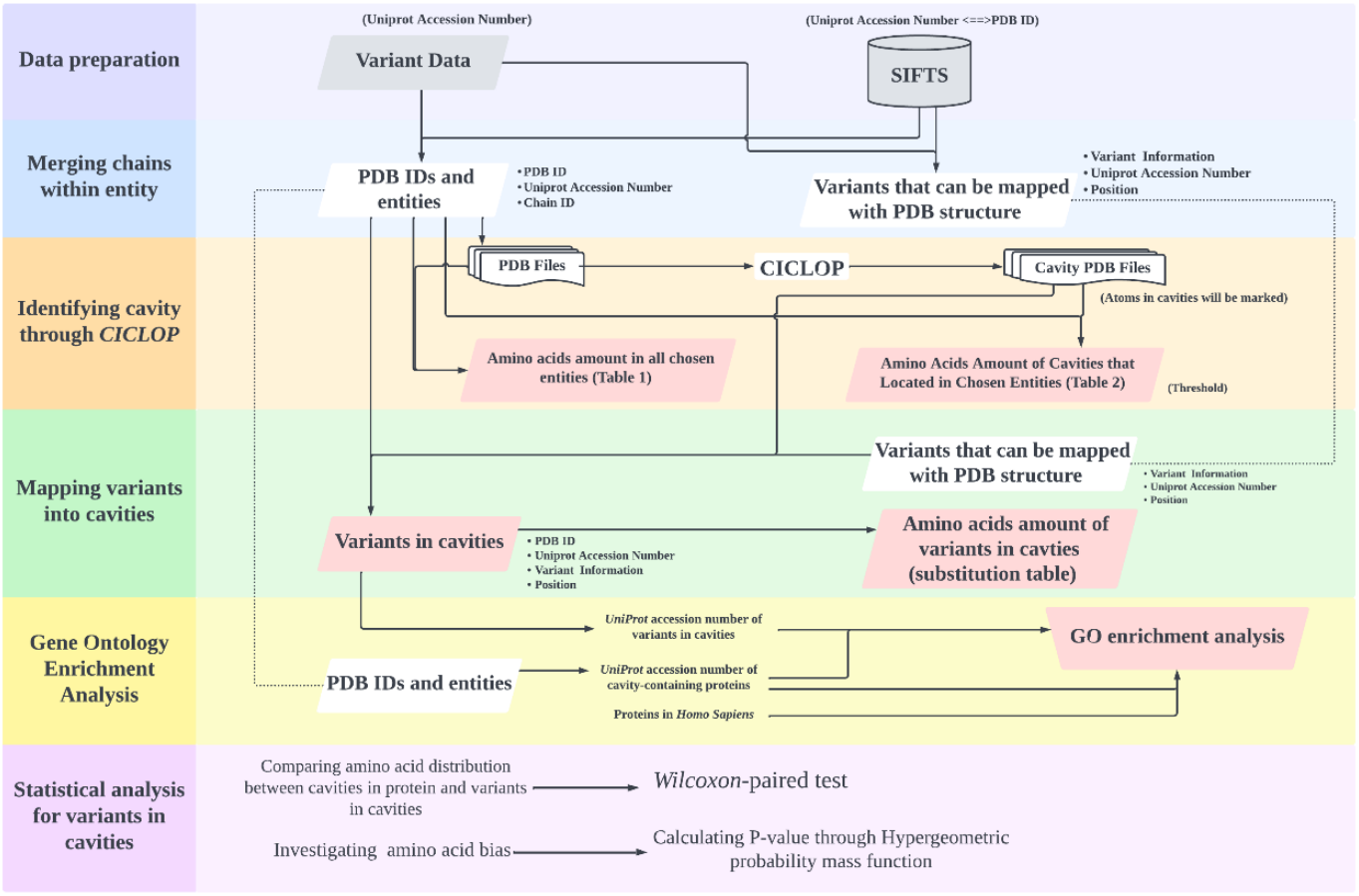

Figure 2. This is the flowchart of this project. The column on the left side is main steps of this project: data preparation is to download the variation dataset and database from *SIFT*; merging chains within entity is to remove the identical chains under same entities; identifying cavities through *CICLOP* is to detect cavities for PDB structure; mapping variants into cavities is to filter all variants that locates in cavities; GO enrichment analysis is two compare the difference of enrichment between all proteins in variation dataset and the proteins whose variants locate in cavities; statistics analysis is to check whether the amino acid distribution of variants in cavities have bias.

The amount of PDB ID, uniport ID and variant decreased after couple of round filtering (Figure 3). In the initial variation data, the number of proteins that have PDB structure is 923, and these 923 proteins include 7 130 amino acid substitution variants. Also, these 923 proteins could be mapped with 33 736 PDB structures, and in *SIFT* database these 33 736 PDB structures have 72 944 chains whose Uniport ID mentioned in variation dataset.

In PDB database, the entities of PDB structure might have couple of chain IDs recorded by SIFT database, and these chains have identical amino acid sequence, so we randomly left one chain for each

entity. Also, in some entities, the chain ID assigned by PDB database is different from the character named by author in publication and both IDs are written in PDB database. The ID assigned by PDB database is used in PDB files, while ID named by author is used in mmCIF file. Because the following steps use PDB files, so in this case we will choose the chain ID assigned by PDB database (Berman et al. 2000) (Dana et al. 2019)

After we remove the identical amino acid chains in same entity of each PDB entities, the total number of PDB chains decreased from 72 944 to 41 639, this is the amount of entity for all chosen PDB structure as well.

Next, we upload all 33 736 PDB files into *CICLOP*, the result shows that 33 317 PDB structures have cavities detected by CICLOP. And among these 33 317 cavity-containing PDB structures, 32 825 PDB structure whose cavities locate in entities selected previously.

In variation data, it includes the position of variant substitution. And when we mapped 7 130 variants into PDB structures, it shows that there are 490 proteins including 4 045 variants locate in cavity areas of PDB structures. And these cavity-located variants can be found in 7 131 entities from 6900 PDB structures.

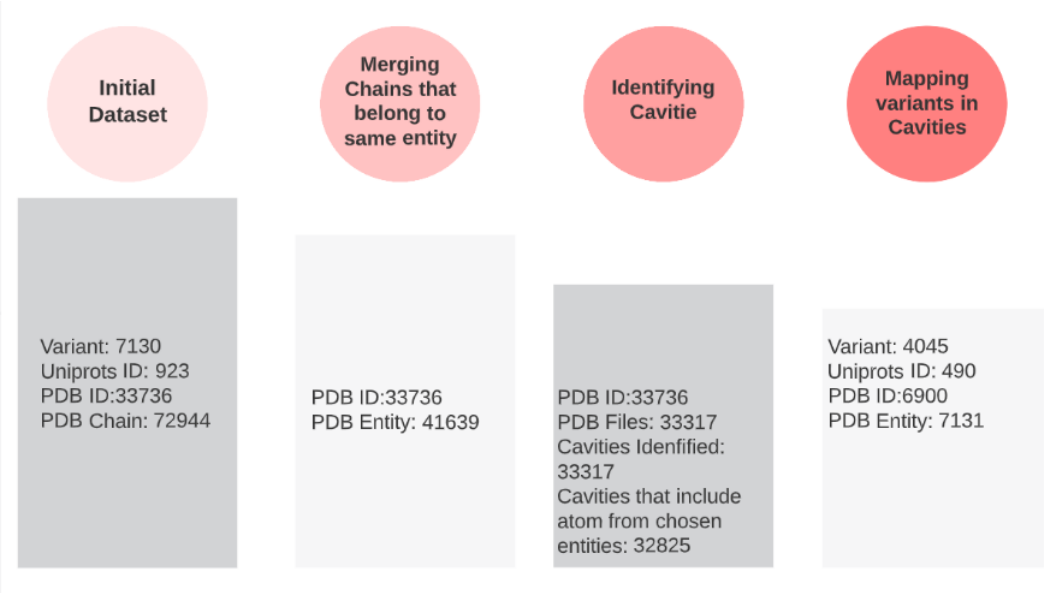

Figure 3. This figure displays the amount of variant, PDB structures, entities and Uniport IDs in different step.

Discovering cavities detected by *CICLOP*

The amino acid distribution of cavities found by *CICLOP*

When we finish the cavities detection, we count and summarize the amino acid distribution of cavities in proteins, and we also count the amino acid distribution of all cavity-detected entities.For the amount of amino acid in cavities, we set a “threshold”, and it is an integer which means the fewest amount of atom that is detected by *CICLOP* should be equal and greater than this integer.

Finally, we make a bar-chart to show the amount of amino acid in cavities under different "threshold". The most common amino acids in chosen entities is *Leucine* which covers 1 million amino acids, while the number of most kinds of amino acids are around 500 000. And it seems like the amount of each kind of amino acids in protein cavities make up around 30% of single type of amino acids in chosen entities. And the bar-chart (Figure 4) shows that the amount of amino acid in cavities doesn't change too much under different "threshold", which could indicate that the depth of detection of *CICLOP* is longer than two or three atoms.

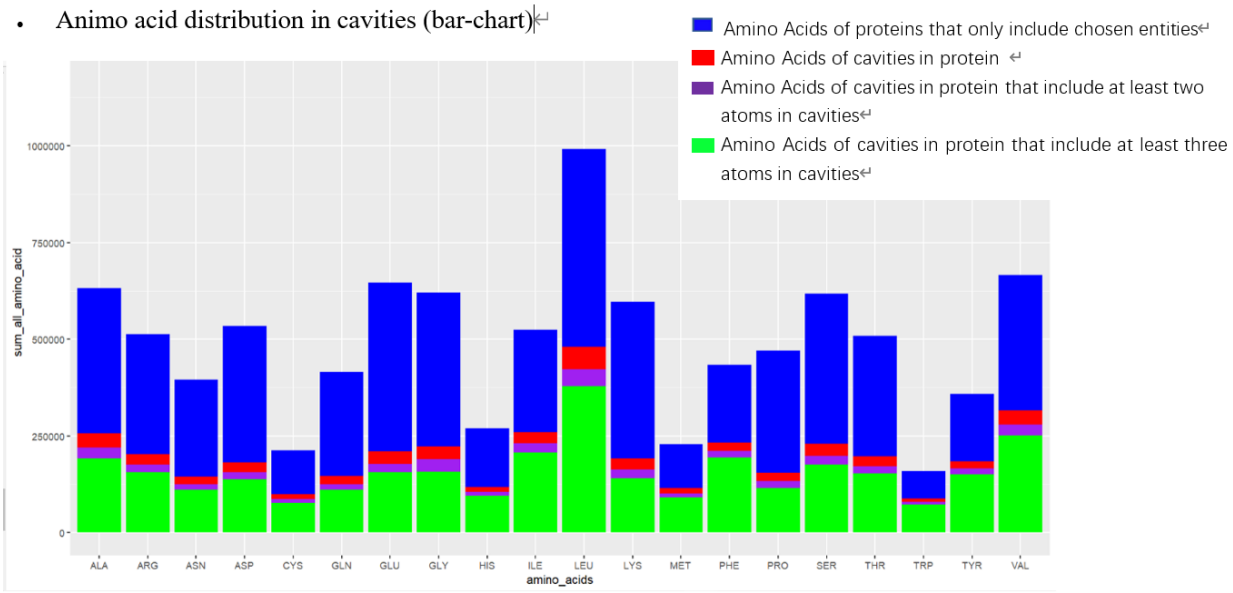

Figure 4. This figure the distribution of amino acids. For every kind of amino acid, it has four bars overlapped together (blue, red, purple and green), which means the length of every bar should be measured from zero. The blue bars indicate the number of amino acids of cavity-located entities; the red bar indicate the amino acids amounts of cavities detected by *CICLOP*; the purple bar indicate the amino acids amount in cavities, and the "threshold" is two, which means that only amino acid that have two atoms detected as cavity can be counted; the green bar indicate the amounts of amino acids, under the "threshold" of three.

Paired t-test* and *Wilcoxon signed rank exact test

To compare the amino acid distribution of cavities detected by *CICLOP* with the distribution of entities mapped from variant proteins, we normalized these two groups of data, and used *paired t-test* and *Wilcoxon signed rank exact test* to check whether two groups are significantly different (Figure 4). In *paired t-test*, the result is "p-value=1"; in *Wilcoxon signed rank exact test,* p-value is 0.9273, and these results (p-value > 0.05) indicates that there might not be significant difference between amino acid distribution of protein cavities and the distribution of chosen protein entities that are mapped from variant proteins.

Discovering variants in cavities

The amino acid distribution of substitution variants located in cavities

When we mapped amino acid substitution into PDB structures which have cavity-located entities, the result shows that there are 4 045 variants found. And the amino distribution of variant indicates that the amount *arginine* is 600 which is much higher than other kinds of amino acid (Figure 5). While the amount of most kinds of amino acids is around 200, such as *Alanine, Serine, Valine.* And there are couple of kind of amino acids that might have lowest frequency of mutation, such as *Asparagine* and *Tryptophan*,

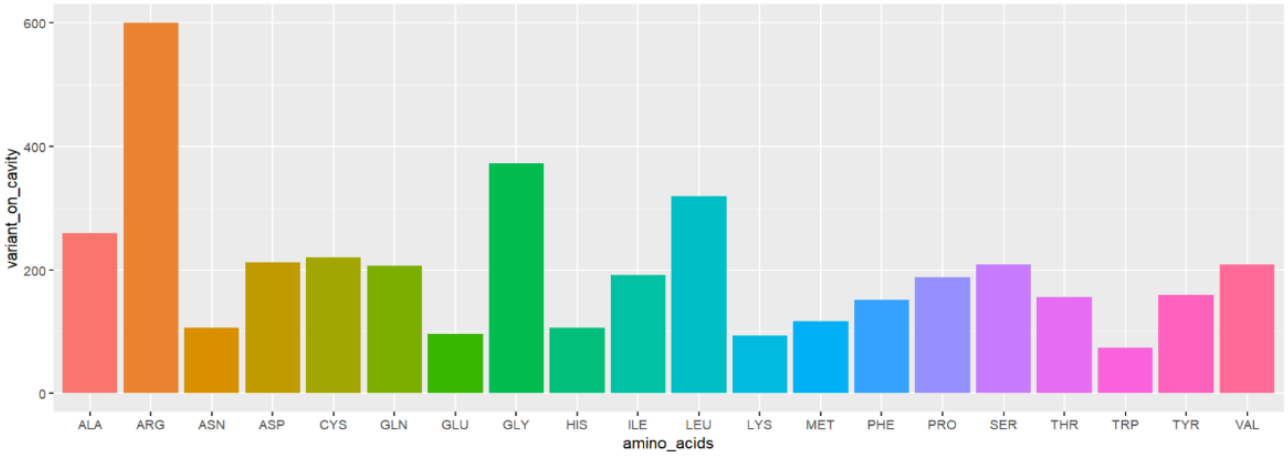

Figure 5. This figure displays the amino acid distribution of substitution variant in cavities.

The bias inside amino acid distribution of substitution variants

To check the amino acid bias of substitution variants on the background of amino acids in cavities, compared with randomly sampling without return, we use hypergeometric function to calculate the fold changed value (Plaisier et al. 2010).

The result can be generated into a table (Table 1). Generally, the p-values of 20 kinds of amino acids are lower than 0.05. The fold change value of most kinds of amino acids are from 1 to 1.5. And the fold change values of *Alanine, Glutamic Acid, Methionine* are around 1, which means the number of actual mutations in cavities might similar to sampling stimulation.

Especially, fold changed value of *arginine* around 3, and it is over-enriched, which indicates that *arginine* has the highest frequency of mutation compared with other kinds of amino acid. Also, the fold changed value of *cysteine* is than 2 and it is over-enriched as well, which means *cysteine* has the second highest frequency of substitution. The reason might be that both *arginine* and *cysteine* have high percentage of GC content, and high mutability of CpG dinucleotides are common in these amino acids, known as mutational hotspots (Ollila, Lappalainen, and Vihinen 1996).

	sum_cavity_amino_acid_1	variant_on_cavity	P_value	Expected_Value	Expect_Text	Fold_Change_Value
A	257211	259	0.0256	258.142674885204	over-enriched	1.00332112896551
R	202387	600	9.26e-122	203.12009028382	over-enriched	2.95391755272272
N	144450	106	9.89e-05	144.97322971089	under-enriched	1.36767197840462
D	181046	212	0.00227	181.701788482089	over-enriched	1.16674690860788
C	99651	220	2.5e-26	100.01195786722	over-enriched	2.19973695837534
Q	146006	96	1.43e-06	146.534865885553	under-enriched	1.52640485297451
E	209596	207	0.0277	210.355202869392	under-enriched	1.01620870951397
G	223115	372	2.27e-21	223.923171664556	over-enriched	1.66128407897539
H	117724	106	0.0202	118.15042225327	under-enriched	1.11462662503085
I	260069	192	6.72e-07	261.011027190595	under-enriched	1.35943243328435
L	480260	319	1.42e-17	481.999607483226	under-enriched	1.51097055637375
K	191702	93	1.79e-16	192.396386860762	under-enriched	2.06877835334152
M	114974	117	0.037	115.390461147662	over-enriched	1.01394863003691
F	232058	151	9.6e-10	232.898565179991	under-enriched	1.542374603841
P	154774	188	0.00104	155.334625512449	over-enriched	1.21029036108201
S	229694	209	0.00944	230.526002251389	under-enriched	1.1029952260832
T	197565	156	0.000198	198.280623937916	under-enriched	1.27102964062767
W	87899	74	0.0135	88.2173895351852	under-enriched	1.19212688561061
Y	183799	159	0.00468	184.464760454357	under-enriched	1.16015572612803
V	316421	209	3.86e-12	317.567146544475	under-enriched	1.51946003131328

Table 1. This table displays the result of fold changed value for each kind of amino acid. The row names are character for twenty amino acids. For the column name, the 1st column indicates the number of amino acids in cavities detected by *CICLOP*; the 2nd column is the amounts of amino acids that are substitution variants in cavities; the 3rd column is p-value of fold change; the 4th column is expected value calculated through hypergeometric function; the 5th column is the direction of enrichment; the6th column is fold changed value.

Discovering the mutation frequency amino acid substitution of variants in cavities

Any kind of amino acids could be substituted by other 19 sorts of amino acid in human substitution mutation. To get the insight of amino acid substitution distribution happened in the same kind of amino acids, we count every kind of mutations that substitute the same amino acid, and calculating the percentage (Table 2).

..	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	0	0	0	0.11	0	0	0.05	0.05	0	0	0	0	0	0	0.14	0.07	0.29	0	0	0.27
R	0	0	0	0	0.2	0.17	0	0.1	0.16	0.01	0.05	0.03	0.01	0	0.07	0.03	0.02	0.16	0	0
N	0	0	0	0.11	0	0	0	0	0.1	0.07	0	0.26	0	0	0	0.29	0.08	0	0.07	0.01
D	0.05	0	0.27	0	0	0	0.08	0.21	0.13	0	0	0	0	0	0	0	0	0	0.12	0.14
C	0	0.22	0	0	0	0	0	0.08	0	0	0	0	0	0.14	0	0.13	0	0.05	0.37	0
Q	0	0.33	0	0	0	0	0.09	0	0.17	0	0.05	0.15	0	0	0.21	0	0	0	0	0
E	0.07	0	0	0.11	0	0.09	0	0.18	0	0	0	0.52	0	0	0	0	0	0	0	0.03
G	0.07	0.27	0	0.14	0.05	0	0.14	0	0	0	0	0	0	0	0	0.17	0	0.01	0	0.15
H	0.01	0.34	0.04	0.08	0	0.1	0	0	0	0	0.08	0	0	0	0.12	0	0	0	0.24	0
I	0	0.01	0.1	0	0	0	0	0	0	0	0.04	0.01	0.09	0.1	0	0.08	0.36	0	0	0.21
L	0.01	0.12	0	0	0	0.06	0	0	0.03	0.01	0	0	0.03	0.15	0.36	0.08	0	0.02	0	0.13
K	0.01	0.27	0.17	0	0	0.09	0.3	0	0	0.03	0	0	0.03	0	0	0	0.1	0	0	0
M	0	0.1	0	0	0	0	0	0	0	0.17	0.08	0.13	0	0	0	0	0.25	0	0	0.27
F	0.01	0	0	0	0.15	0	0	0	0	0.07	0.37	0	0	0	0	0.23	0	0	0.05	0.13
P	0.06	0.1	0	0	0	0.02	0	0	0.05	0	0.42	0	0	0	0	0.24	0.11	0	0	0
S	0.03	0.1	0.08	0	0.07	0	0	0.06	0	0.05	0.17	0	0	0.12	0.19	0	0.03	0.02	0.07	0
T	0.19	0.05	0.04	0	0	0	0.01	0	0	0.3	0	0.03	0.19	0	0.11	0.08	0	0	0	0
W	0	0.34	0	0	0.34	0	0	0.12	0	0	0.08	0	0	0	0	0.12	0	0	0	0
Y	0.01	0	0.08	0.09	0.43	0	0	0.01	0.21	0	0	0	0	0.06	0	0.1	0	0.01	0	0
V	0.15	0	0	0.05	0	0	0.06	0.08	0	0.19	0.11	0	0.25	0.11	0	0	0	0	0	0
sum_row	0.04	0.09	0.03	0.04	0.07	0.04	0.03	0.05	0.05	0.04	0.06	0.05	0.03	0.04	0.07	0.08	0.06	0.03	0.04	0.07

Table 2. This table is the percentage of different kinds of amino acid substitution that happened in same amino acids. The twenty rows in blue indicate the original amino acid in sequence before substitution, and the final row "sum_row" indicate the percentage of variant amino acid after substitution. The columns mean the amino acids after substitution.

The numbers in red indicate the percentage in range 0.25-0.40, while the numbers in green are percentages higher than 0.40.

The result show that the distribution of variant amino acids after substitution seems average, which means the percentage of each kind of post-substitution amino acids all lower than 0.10. However, when analyzing 20 kinds of amino acid in human separately, we could find that the distribution of 19 amino acids that substitute same original ones is not average. More specifically, for each kind of amino acid that have mutation, around 25% or higher of these amino acids would be substituted by one kind of amino acids. And some of substitution pairs could occupy over 40% within all amino acids that substitute the same amino acid. There are three in our sample: 52% of glutamic acid substituted by lysine; 43% of proline substituted by leucine; 43% of tyrosine substituted by cysteine.

Gene ontology enrichment analysis

To discover the whether the biological function of proteins that have variants in cavities could be enriched in some GO terms, we need compare the GO annotation. We add GO annotation for variants proteins, and the background of annotation is human proteome. We also add GO annotation for the proteins that have substitution variants in cavities found by *CICLOP*, and the background is all proteins in sample.

We annotate three types of GO terms (Biological Processes, Molecular Functions, Cellular Components) to proteins and get the table ranked by p-value (Tables S1-S6).

The cellular component of GO annotation (Tables S1 and S2) show that the GO terms of both groups might be related to lumen structure for transportation of secreted proteins. The molecular function of GO annotation (Tables S3 and S4) indicate that the GO terms of both groups seem mainly related to binding, such as binding active sites of enzymes or binding zymolyte in enzyme catalysis. And the biological process of GO annotation (Tables S5 and S6) indicate that the GO terms of both groups seem mainly related to signal regulation in immune reaction such as T cell activation.

We find that there might be no much difference between the enrichment of all proteins in sample and the proteins that include substitution mutation in cavities detected by *CICLOP.* It might because the number of proteins that have substitution in cavities is 490 which makes up 53% of all proteins in sample (923 proteins). Maybe the GO term distribution of 490 proteins that have variants in cavities is similar to all proteins' distribution in sample.

Discussion

In the project, we use experimentally validated amino acid substitution data in human as sample to discover the regulation in protein cavities area. We firstly use *SIFT* to map uniport ID with PDB structure, and then remove the identical amino acids chains under same entity. Then using CICLOP to

detect cavities in these PDB structures and only leave the cavity information that located in the entities whose uniport ID mentioned in sample. Next, we do a series of statistical analysis to compare all PDB structures mapped by *SIFT* and protein cavity structures; and compare the substitution amino acids in cavities and amino acids in protein cavities found by *CICLOP*.

We found that the amino acid distribution of cavities detected by *CICLOP* is not significantly different from the distribution of all PDB entities whose uniport ID is the same as the uniport ID in sample. And most of amino acids that locate in protein cavities have at least 2 or 3 atoms detected by *CICLOP*. When we map the substitution mutation to PDB structures and get the variants that locate in cavities found by *CICLOP*. For the amino acids that are substituted, we found the amino acids whose codons have higher GC content, for instance arginine, could be substituted by other amino acid more frequently. However, after substitution, the distribution of amino acids which substitute original ones seems average. When we observe the substitution of every kind of amino acids, we could find the percentage of 19 kinds of amino acids that substitute the same amino acids might have bias, which means among these 19 amino acids, there would be at least one sort of amino acid that might occupy around 30% (or even more) of whole substitution occurs in one original amino acid.

The short coming of this project is that it doesn't use different protein cavities detecting method to process sample of this project, and compare the result of different tools, checking whether *CICLOP* surpass the limitation of previous tools. And the data analysis is all basic statistical test, and it doesn't include more analysis in the area of chemistry such as hydropathy level, because this chemical feature might determine the function of cavity regions (Chwastyk et al. 2020).

In this project, there are approximately half of the variant proteins have mutations on cavity surfaces. And GO annotation of these half amounts of variant proteins might be similar to the GO annotation of all variant proteins in sample, which means the GO terms of proteins that have mutation on cavities might not be enriched into a limited scale of GO terms. Maybe, the GO terms would be enriched better, if the number of proteins in sample increased dramatically.

The variants that locate on the surface of protein cavities might be special, because the structure of cavities has significant biological function such as binding, signal reception and so on. The variants in protein cavities might change the micro-environment around cavities and influence the stability of protein, but it still need a lot of deep research.

Acknowledge

Thank Mauno for giving a chance to do this research project and guiding me patiently. This is a tough project, but we got some result somehow finally. And thank Anna to examine my project and give me feedback. Finally, I should thank myself a lot as well.

Reference

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." *Nature Genetics* 25(1):25–29. doi: 10.1038/75556.

Bateman, Alex, Maria Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, Emily H. Bowler-Barnett, Ramona Britto, Hema Bye-A-Jee, Austra Cukura, Paul Denny, Tunca Dogan, Thank God Ebenezer, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Marija Lugaric, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Alok Mishra, Katie Moulang, Andrew Nightingale, Sangya Pundir, Guoying Qi, Shriya Raj, Pedro Raposo, Daniel L. Rice, Rabie Saidi, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan J. Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H. Auchincloss, Kristian B. Axelsen, Parit Bansal, Delphine Baratin, Teresa M. Batista Neto, Marie Claude Blatter, Jerven T. Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard de Castro, Anne Estreicher, Maria L. Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Venkatesh Muthukrishnan, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J. A. Sigrist, Karin Sonesson, Shyamala Sundaram, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, and Jian Zhang. 2023. "UniProt: The Universal Protein Knowledgebase in 2023." *Nucleic Acids Research* 51(D1):D523–31. doi: 10.1093/NAR/GKAC1052.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28(1):235–42. doi: 10.1093/NAR/28.1.235.

Choi, Umji, and Chang Ro Lee. 2019. "Distinct Roles of Outer Membrane Porins in Antibiotic Resistance and Membrane Integrity in Escherichia Coli." *Frontiers in Microbiology* 10(APR). doi: 10.3389/FMICB.2019.00953.

Chwastyk, Mateusz, Ewa A. Panek, Jan Malinowski, Mariusz Jaskólski, and Marek Cieplak. 2020. "Properties of Cavities in Biological Structures—A Survey of the Protein Data Bank." *Frontiers in Molecular Biosciences* 7:314. doi: 10.3389/FMOLB.2020.591381/BIBTEX.

Dana, Jose M., Aleksandras Gutmanas, Nidhi Tyagi, Guoying Qi, Claire O'Donovan, Maria Martin, and Sameer Velankar. 2019. "SIFTS: Updated Structure Integration with Function, Taxonomy

and Sequences Resource Allows 40-Fold Increase in Coverage of Structure-Based Annotations for Proteins." *Nucleic Acids Research* 47(D1):D482–89. doi: 10.1093/NAR/GKY1114.

Garg, Parth, Sukriti Sacher, Prutyay Gautam, and Arjun Ray. n.d. "CICLOP: A Robust, Faster, and Accurate Computational Framework for Protein Inner Cavity Detection." doi: 10.1101/2020.11.25.399246.

Ollila, Juha, Ilkka Lappalainen, and Mauno Vihinen. 1996. "Sequence Specificity in CpG Mutation Hotspots." *FEBS Letters* 396(2–3):119–22. doi: 10.1016/0014-5793(96)01075-7.

Plaisier, Seema B., Richard Taschereau, Justin A. Wong, and Thomas G. Graeber. 2010. "Rank–Rank Hypergeometric Overlap: Identification of Statistically Significant Overlap between Gene-Expression Signatures." *Nucleic Acids Research* 38(17):e169–e169. doi: 10.1093/NAR/GKQ636.

Pravda, Lukáš, David Sehnal, Dominik Toušek, Veronika Navrátilová, Václav Bazgier, Karel Berka, Radka Svobodová Vařeková, Jaroslav Koča, and Michal Otyepka. 2018. "MOLEonline: A Web-Based Tool for Analyzing Channels, Tunnels and Pores (2018 Update)." *Nucleic Acids Research* 46(W1):W368–73. doi: 10.1093/NAR/GKY309.

Sasidharan Nair, Preethy, and Mauno Vihinen. 2013. "VariBench: A Benchmark Database for Variations." *Human Mutation* 34(1):42–49. doi: 10.1002/HUMU.22204.

Tanwar, Ajay Singh, Venuka Durani Goyal, Deepanshu Choudhary, Santosh Panjikar, and Ruchi Anand. 2013. "Importance of Hydrophobic Cavities in Allosteric Regulation of Formylglycinamide Synthetase: Insight from Xenon Trapping and Statistical Coupling Analysis." *PloS One* 8(11). doi: 10.1371/JOURNAL.PONE.0077781.

Xu, Youjun, Shiwei Wang, Qiwan Hu, Shuaishi Gao, Xiaomin Ma, Weilin Zhang, Yihang Shen, Fangjin Chen, Luhua Lai, and Jianfeng Pei. 2018. "CavityPlus: A Web Server for Protein Cavity Detection with Pharmacophore Modelling, Allosteric Site Identification and Covalent Ligand Binding Ability Prediction." *Nucleic Acids Research* 46(W1):W374–79. doi: 10.1093/NAR/GKY380.

Yang, Yang, Aibin Shao, and Mauno Vihinen. 2022. "PON-All: Amino Acid Substitution Tolerance Predictor for All Organisms." *Frontiers in Molecular Biosciences* 9:529. doi: 10.3389/FMOLB.2022.867572/BIBTEX.

Zhou, Huan Xiang, and J. Andrew McCammon. 2010. "The Gates of Ion Channels and Enzymes." *Trends in Biochemical Sciences* 35(3):179–85. doi: 10.1016/J.TIBS.2009.10.007.

Supplementary Material

The amino acid substitution dataset in this project can be found at: http://structure.bmc.lu.se/VariBench/EXTENSION/VariationTypeDatasets/SubstitutionsCodingRegion/TrainingDatasets/Dataset25/All_species_train.csv

The *SIFT* database that used in this project can be found at: ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/flatfiles/tsv/Uniport_segments_observed.tsv.gz

Table S1. This table is cellular component GO annotation for all proteins in sample. Here is the top 15 GO term ranked by p-value.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0005759	GO:0005759	mitochondrial matrix	204/3926	480/19550	9.159035e-30	6.722731e-27	4.010693e-27	8050/5138/5442/4706/7015/9692/587/4968/2309/10469/99...	204
GO:0031983	GO:0031983	vesicle lumen	148/3926	327/19550	3.236765e-25	1.187893e-22	7.086813e-23	5768/6386/1654/2219/8566/317/5709/81/9535/8993/3417/...	148
GO:0060205	GO:0060205	cytoplasmic vesicle lumen	147/3926	325/19550	5.095769e-25	1.246765e-22	7.438035e-23	5768/6386/1654/2219/8566/317/5709/81/9535/8993/3417/...	147
GO:0034774	GO:0034774	secretory granule lumen	145/3926	322/19550	1.849125e-24	3.393145e-22	2.024305e-22	5768/6386/1654/2219/8566/317/5709/81/9535/8993/3417/...	145
GO:0031252	GO:0031252	cell leading edge	168/3926	422/19550	4.530039e-21	6.650097e-19	3.967360e-19	4641/7099/322/1654/3636/9051/10253/23396/1500/23380/...	168
GO:0045121	GO:0045121	membrane raft	139/3926	335/19550	1.609500e-19	1.687676e-17	1.006845e-17	4641/5348/8797/28514/6386/1200/3551/8871/4864/7097/3...	139
GO:0098857	GO:0098857	membrane microdomain	139/3926	335/19550	1.609500e-19	1.687676e-17	1.006845e-17	4641/5348/8797/28514/6386/1200/3551/8871/4864/7097/3...	139
GO:0005775	GO:0005775	vacuolar lumen	86/3926	174/19550	3.754205e-18	3.444483e-16	2.054933e-16	6386/1200/8029/1515/1727/12/718/7276/2638/2512/383/2...	86
GO:0030055	GO:0030055	cell-substrate junction	158/3926	425/19550	1.293271e-16	1.054734e-14	6.292404e-15	129446/6386/8573/4659/2580/247/726/81/23396/4868/101...	158
GO:0005925	GO:0005925	focal adhesion	155/3926	418/19550	3.289105e-16	2.414203e-14	1.440282e-14	129446/6386/8573/4659/2580/247/726/81/23396/4868/101...	155
GO:0009897	GO:0009897	external side of plasma membrane	154/3926	421/19550	1.626905e-15	1.085589e-13	6.476482e-14	7099/11119/2219/4065/8029/1515/23308/10159/19/2936/2...	154
GO:0045177	GO:0045177	apical part of cell	157/3926	435/19550	3.341015e-15	2.043587e-13	1.219177e-13	5348/28514/9854/1856/8972/5205/8029/1515/9026/10159/...	157
GO:0030139	GO:0030139	endocytic vesicle	129/3926	336/19550	4.903626e-15	2.768663e-13	1.651748e-13	4641/7476/1856/9146/6892/10618/8029/7097/23380/9727/...	129
GO:0030016	GO:0030016	myofibril	98/3926	231/19550	6.670755e-15	3.497381e-13	2.086492e-13	129446/8789/845/4659/23336/8557/81/9997/11155/23363/...	98
GO:0043202	GO:0043202	lysosomal lumen	54/3926	97/19550	1.037748e-14	5.078047e-13	3.029496e-13	1200/8029/1515/2638/2717/3073/1509/5660/3074/1514/15...	54

Table S2. This table is cellular component GO annotation for proteins which include variant-located cavities. Here is the top 15 GO term ranked by p-value.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0030141	GO:0030141	secretory granule	76/477	311/3926	2.691726e-10	1.092841e-07	7.876841e-08	1654/317/3417/6647/2157/2162/4860/1675/5340/5265/718...	76
GO:0099503	GO:0099503	secretory vesicle	80/477	353/3926	4.199622e-09	5.920551e-07	4.267340e-07	10059/1654/317/3417/6647/2157/2162/4860/1675/5340/52...	80
GO:0031983	GO:0031983	vesicle lumen	44/477	148/3926	4.374791e-09	5.920551e-07	4.267340e-07	1654/317/3417/2157/2162/4860/1956/1675/5340/5265/718...	44
GO:0034774	GO:0034774	secretory granule lumen	43/477	145/3926	7.227006e-09	7.335411e-07	5.287125e-07	1654/317/3417/2157/2162/4860/1675/5340/5265/718/7040...	43
GO:0009986	GO:0009986	cell surface	76/477	334/3926	9.141408e-09	7.422823e-07	5.350129e-07	7099/11119/4040/2936/1956/2147/5340/718/3949/7040/92...	76
GO:0060205	GO:0060205	cytoplasmic vesicle lumen	43/477	147/3926	1.144184e-08	7.742312e-07	5.580406e-07	1654/317/3417/2157/2162/4860/1675/5340/5265/718/7040...	43
GO:1904813	GO:1904813	ficolin-1-rich granule lumen	21/477	53/3926	2.927429e-07	1.697909e-05	1.223797e-05	1654/317/3417/4860/1675/5265/1471/2934/1508/4318/272...	21
GO:0005783	GO:0005783	endoplasmic reticulum	95/477	495/3926	8.531733e-07	4.329854e-05	3.120818e-05	6820/4221/10059/27429/5071/8996/4040/2157/1956/2147/...	95
GO:0043235	GO:0043235	receptor complex	45/477	182/3926	1.204558e-06	5.433894e-05	3.916574e-05	7099/3551/5364/27429/1956/3949/2688/920/5284/7037/70...	45
GO:0101002	GO:0101002	ficolin-1-rich granule	25/477	79/3926	3.127636e-06	1.269820e-04	9.152451e-05	1654/317/3417/4860/1675/5265/1471/3689/2934/1508/160...	25
GO:0098552	GO:0098552	side of membrane	51/477	228/3926	5.699959e-06	2.103803e-04	1.516353e-04	7099/11119/3551/27429/2936/2147/5340/3845/3949/920/3...	51
GO:0030134	GO:0030134	COPII-coated ER to Golgi transport vesicle	14/477	32/3926	7.443084e-06	2.518243e-04	1.815068e-04	2157/5265/3106/3117/3123/3119/3105/3115/351/3133/311...	14
GO:0005739	GO:0005739	mitochondrion	89/477	479/3926	8.583356e-06	2.680648e-04	1.932124e-04	10059/32/8473/587/4968/27429/5071/8996/3417/2744/913...	89
GO:0072562	GO:0072562	blood microparticle	20/477	59/3926	9.680752e-06	2.807418e-04	2.023495e-04	2162/2147/5340/462/718/7040/335/2244/6521/325/2335/2...	20
GO:0098576	GO:0098576	lumenal side of membrane	9/477	15/3926	1.367371e-05	3.401434e-04	2.451643e-04	3106/3117/3123/3119/3105/3115/5476/3133/3113	9

Table S3. This table is molecular function GO annotation for all proteins in sample. Here is the top 15 GO term ranked by p-value.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0033218	GO:0033218	amide binding	157/3916	400/18368	1.414394e-16	1.707173e-13	9.781649e-14	7099/322/773/32/9854/9536/1200/6892/8811/3061/7097/1...	157
GO:0004713	GO:0004713	protein tyrosine kinase activity	71/3916	136/18368	1.943116e-15	1.172670e-12	6.719089e-13	2051/8767/8444/3717/8576/25/1956/2064/4914/3643/3932...	71
GO:0019199	GO:0019199	transmembrane receptor protein kinase activity	73/3916	143/18368	3.704927e-15	1.490615e-12	8.540831e-13	2051/8767/8444/3717/8576/25/1956/2064/4914/3643/3932...	73
GO:0004674	GO:0004674	protein serine/threonine kinase activity	148/3916	386/18368	1.063372e-14	3.208726e-12	1.838515e-12	6446/29904/10733/3551/8573/2580/375449/3656/8767/84...	148
GO:1901681	GO:1901681	sulfur compound binding	111/3916	265/18368	2.470435e-14	5.963629e-12	3.417001e-12	32/9536/2660/2255/10563/9731/8435/80739/30008/2147/4...	111
GO:0042277	GO:0042277	peptide binding	126/3916	321/18368	1.433428e-13	2.883580e-11	1.652215e-11	7099/322/773/9854/9536/1200/6892/8811/3061/7097/1026...	126
GO:0004714	GO:0004714	transmembrane receptor protein tyrosine kinase activity	63/3916	124/18368	3.718939e-13	6.412513e-11	3.674200e-11	2051/8767/8444/3717/8576/25/1956/2064/4914/3643/3932...	63
GO:0051287	GO:0051287	NAD binding	37/3916	56/18368	6.815314e-13	1.028261e-10	5.891660e-11	26227/3420/7358/4720/3417/3939/216/2746/1727/217/127...	37
GO:0140272	GO:0140272	exogenous protein binding	45/3916	77/18368	1.443947e-12	1.936494e-10	1.109559e-10	4864/1956/3949/920/2993/213/7037/3690/3383/1604/3678...	45
GO:0001618	GO:0001618	virus receptor activity	44/3916	76/18368	4.006579e-12	4.720976e-10	2.704994e-10	4864/1956/3949/920/2993/7037/3690/3383/1604/3678/468...	44
GO:0005543	GO:0005543	phospholipid binding	163/3916	466/18368	4.302463e-12	4.720976e-10	2.704994e-10	9743/3092/7287/5286/6861/2494/889/6386/9854/1200/533...	163
GO:0002020	GO:0002020	protease binding	64/3916	135/18368	1.283006e-11	1.290491e-09	7.394168e-10	8797/8767/5071/8996/9507/8915/462/5265/1471/3949/127...	64
GO:0140030	GO:0140030	modification-dependent protein binding	72/3916	160/18368	1.565065e-11	1.453103e-09	8.325894e-10	8805/23378/23476/5253/9682/100137047/3329/7409/5336...	72
GO:0019842	GO:0019842	vitamin binding	68/3916	148/18368	1.789015e-11	1.542387e-09	8.837467e-10	32/8566/5264/10558/9517/8974/8029/8985/8879/2806/213...	68
GO:0045296	GO:0045296	cadherin binding	123/3916	332/18368	2.792490e-11	2.247024e-09	1.287485e-09	1192/10528/1654/11122/9973/1500/8615/2317/3417/9217/...	123

Table S4. This table is molecular function GO annotation for proteins which include variant-located cavities. Here is the top 15 GO term ranked by p-value.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0002020	GO:0002020	protease binding	23/477	64/3916	6.624210e-07	0.0002353355	0.0001929199	8767/5071/8996/462/5265/1471/3949/1281/2335/5621/745...	23
GO:0051087	GO:0051087	chaperone binding	17/477	39/3916	8.465306e-07	0.0002353355	0.0001929199	332/5071/6647/5340/3106/2244/2335/213/5621/7450/7157...	17
GO:0004175	GO:0004175	endopeptidase activity	36/477	132/3916	1.343886e-06	0.0002490668	0.0002041763	10753/27429/2147/1675/5340/2161/5972/4057/2160/354/1...	36
GO:0008233	GO:0008233	peptidase activity	45/477	185/3916	2.103543e-06	0.0002923925	0.0002396932	10753/27429/2147/1675/5340/2161/5972/7037/4057/2160/...	45
GO:0005102	GO:0005102	signaling receptor binding	91/477	483/3916	3.993531e-06	0.0004440807	0.0003640419	7099/11119/2660/3551/8573/5364/8767/5071/9046/1500/8...	91
GO:0030234	GO:0030234	enzyme regulator activity	73/477	370/3916	8.047679e-06	0.0007457516	0.0006113412	10059/1654/317/332/5364/8996/4040/8412/216/1956/462/...	73
GO:0046914	GO:0046914	transition metal ion binding	73/477	372/3916	9.858442e-06	0.0007830420	0.0006419106	4194/5071/8856/125/5053/6647/2157/760/920/213/7018/4...	73
GO:0061134	GO:0061134	peptidase regulator activity	24/477	79/3916	1.138323e-05	0.0007911347	0.0006485447	317/332/8996/462/5265/718/727/1471/2335/4057/5621/35...	24
GO:0017171	GO:0017171	serine hydrolase activity	22/477	70/3916	1.453653e-05	0.0008980347	0.0007361776	27429/2147/1675/5340/2161/4057/2160/354/1991/2155/43...	22
GO:0044877	GO:0044877	protein-containing complex binding	88/477	480/3916	1.888747e-05	0.0010501433	0.0008608709	10059/1654/5071/1956/4893/3265/3845/3949/920/3117/31...	88
GO:0042277	GO:0042277	peptide binding	32/477	126/3916	2.643709e-05	0.0013362748	0.0010954316	7099/1471/3949/3106/3117/3123/3119/1410/335/5621/310...	32
GO:0016491	GO:0016491	oxidoreductase activity	53/477	255/3916	3.725823e-05	0.0015574558	0.0012767481	23028/3417/9131/125/3939/216/2936/5053/6647/2157/217...	53
GO:0033218	GO:0033218	amide binding	37/477	157/3916	3.766200e-05	0.0015574558	0.0012767481	7099/32/1471/3949/3106/3117/3123/3119/1410/335/5621/...	37
GO:0008236	GO:0008236	serine-type peptidase activity	21/477	69/3916	3.921651e-05	0.0015574558	0.0012767481	27429/2147/1675/5340/2161/4057/2160/354/1991/2155/43...	21
GO:0016903	GO:0016903	oxidoreductase activity, acting on the aldehyde or oxo grou...	11/477	24/3916	4.406919e-05	0.0016334982	0.0013390850	216/217/1738/593/231/594/8659/218/8644/501/316	11

Table S5. This table is biological process GO annotation for all proteins in sample. Here is the top 15 GO term ranked by p-value.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:0018108	GO:0018108	peptidyl-tyrosine phosphorylation	177/3871	375/18723	3.941144e-31	1.497624e-27	6.727336e-28	1856/2051/2255/8767/8444/9467/3717/8576/11116/25/195...	177
GO:0042060	GO:0042060	wound healing	192/3871	422/18723	4.739317e-31	1.497624e-27	6.727336e-28	7099/1192/7476/8573/2255/3717/8324/8291/80739/2157/2...	192
GO:0018212	GO:0018212	peptidyl-tyrosine modification	177/3871	378/18723	1.352789e-30	2.849875e-27	1.280165e-27	1856/2051/2255/8767/8444/9467/3717/8576/11116/25/195...	177
GO:0001819	GO:0001819	positive regulation of cytokine production	203/3871	467/18723	1.774962e-29	2.804440e-26	1.259756e-26	3965/7099/5293/11119/1654/2219/247/8767/7305/7097/37...	203
GO:0042110	GO:0042110	T cell activation	208/3871	487/18723	6.334890e-29	8.007301e-26	3.596884e-26	3965/10326/5293/11119/8943/8600/4092/2175/9092/8767/...	208
GO:0050878	GO:0050878	regulation of body fluid levels	174/3871	379/18723	8.229359e-29	8.668258e-26	3.893786e-26	7099/1192/8835/2255/3783/3717/80739/2157/2162/2806/1...	174
GO:0033674	GO:0033674	positive regulation of kinase activity	200/3871	467/18723	4.945575e-28	4.465148e-25	2.005747e-25	7099/8797/1654/1856/8795/8600/8312/2051/10253/904/37...	200
GO:0002697	GO:0002697	regulation of immune effector process	158/3871	339/18723	3.767826e-27	2.976583e-24	1.337083e-24	3965/7099/117157/259197/4092/8767/7305/4068/8993/87...	158
GO:0007599	GO:0007599	hemostasis	117/3871	222/18723	3.273820e-26	2.298949e-23	1.032689e-23	7099/1192/3717/80739/2157/2162/2147/2158/2159/5340/2...	117
GO:0007596	GO:0007596	blood coagulation	115/3871	217/18723	4.633826e-26	2.928578e-23	1.315519e-23	7099/1192/3717/80739/2157/2162/2147/2158/2159/5340/2...	115
GO:0050817	GO:0050817	coagulation	116/3871	222/18723	1.426775e-25	8.197469e-23	3.682307e-23	7099/1192/3717/80739/2157/2162/2147/2158/2159/5340/2...	116
GO:0070663	GO:0070663	regulation of leukocyte proliferation	122/3871	245/18723	2.170613e-24	1.143190e-21	5.135214e-22	3965/7099/23495/2255/8767/7305/23308/56253/90865/48...	122
GO:0050863	GO:0050863	regulation of T cell activation	149/3871	329/18723	4.299584e-24	2.090260e-21	9.389457e-22	3965/10326/8943/8600/4092/2175/9092/8767/23308/639/1...	149
GO:0070661	GO:0070661	leukocyte proliferation	145/3871	318/18723	7.737502e-24	3.492929e-21	1.569026e-21	3965/7099/11119/8600/23495/2255/8767/7305/23308/101...	145
GO:0022407	GO:0022407	regulation of cell-cell adhesion	185/3871	448/18723	9.133266e-24	3.848150e-21	1.728591e-21	3965/10326/8943/8600/4092/9092/10653/8767/10563/371...	185

Table S6. This table is biological process GO annotation for proteins which include variant-located cavities. Here is the top 15 GO term ranked by p-value.

	ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	qvalue	geneID	Count
GO:1901698	GO:1901698	response to nitrogen compound	96/476	413/3871	3.265537e-11	4.980753e-08	3.384846e-08	7099/4221/2660/8473/4968/8767/5071/23028/4040/8412/9...	96
GO:0043069	GO:0043069	negative regulation of programmed cell death	81/476	325/3871	3.827310e-11	4.980753e-08	3.384846e-08	1654/4194/332/4968/6899/27429/5071/23028/8996/4040/1...	81
GO:0010243	GO:0010243	response to organonitrogen compound	91/476	384/3871	3.885143e-11	4.980753e-08	3.384846e-08	7099/4221/2660/8473/4968/8767/5071/23028/4040/8412/9...	91
GO:1901701	GO:1901701	cellular response to oxygen-containing compound	99/476	438/3871	8.607583e-11	8.276191e-08	5.624376e-08	7099/4221/10059/2660/8473/8767/6899/27429/5071/2302...	99
GO:0052548	GO:0052548	regulation of endopeptidase activity	49/476	157/3871	1.212397e-10	9.325758e-08	6.337646e-08	1654/317/332/8767/27429/8996/9131/462/5265/718/727/1...	49
GO:0043066	GO:0043066	negative regulation of apoptotic process	78/476	317/3871	1.866668e-10	1.073088e-07	7.292545e-08	1654/4194/332/4968/6899/27429/5071/23028/8996/4040/1...	78
GO:0001775	GO:0001775	cell activation	94/476	413/3871	1.953098e-10	1.073088e-07	7.292545e-08	7099/5293/11119/8767/6647/4860/3251/1956/2147/3949/7...	94
GO:0052547	GO:0052547	regulation of peptidase activity	50/476	165/3871	2.527780e-10	1.215230e-07	8.258523e-08	1654/317/332/8767/27429/8996/9131/462/5265/718/727/1...	50
GO:0009410	GO:0009410	response to xenobiotic stimulus	54/476	188/3871	4.310421e-10	1.841987e-07	1.251787e-07	32/4968/8856/8412/11200/3939/6647/4860/1956/5972/335...	54
GO:0010035	GO:0010035	response to inorganic substance	62/476	233/3871	6.295561e-10	2.241076e-07	1.523002e-07	4968/5071/9131/3939/6647/1956/760/1410/2244/7018/562...	62
GO:0042325	GO:0042325	regulation of phosphorylation	103/476	478/3871	6.488266e-10	2.241076e-07	1.523002e-07	7099/4221/1654/2660/3551/8473/8767/6899/5071/4040/84...	103
GO:0014070	GO:0014070	response to organic cyclic compound	85/476	367/3871	6.992437e-10	2.241076e-07	1.523002e-07	10059/32/2660/4968/8767/5071/23028/4040/3417/9131/11...	85
GO:0009611	GO:0009611	response to wounding	62/476	234/3871	7.614458e-10	2.252708e-07	1.530907e-07	7099/8573/8996/6647/2157/2162/1956/2147/5340/2161/46...	62
GO:0002460	GO:0002460	adaptive immune response based on somatic recombinatio...	44/476	140/3871	8.989771e-10	2.469619e-07	1.678316e-07	7099/8767/3251/718/727/3265/7040/920/3106/3117/3123/...	44
GO:1902533	GO:1902533	positive regulation of intracellular signal transduction	79/476	334/3871	1.096815e-09	2.812233e-07	1.911151e-07	7099/5293/10059/1654/3551/5364/8767/6899/5071/8412/6...	79